Advice on Statistical Analysis for Circulation Research

Hideo Kusuoka and Julien I.E. Hoffman

Circulation Research 2002, 91:662-671
doi: 10.1161/01.RES.0000037427.73184.C1
Circulation Research is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0009-7330. Online ISSN: 1524-4571

Advice on Statistical Analysis for Circulation Research

Hideo Kusuoka, Julien I.E. Hoffman

Abstract—Since the late 1970s when many journals published articles warning about the misuse of statistical methods in the analysis of data, researchers have become more careful about statistical analysis, but errors including low statistical power and inadequate analysis of repeated-measurement studies are still prevalent. In this review, several statistical methods are introduced that are not always familiar to basic and clinical cardiologists but may be useful for revealing the correct answer from the data. The aim of this review is not only to draw the attention of investigators to these tests but also to stress the conditions in which they are applicable. These methods are now generally available in statistical program packages. Researchers need not know how to calculate the statistics from the data but are required to select the correct method from the menu and interpret the statistical results accurately. With the choice of appropriate statistical programs, the issue is no longer how to do the test but when to do it. **(*Circ Res*. 2002;91:662-671.)**

Key Words: power analysis ■ repeated measures ■ analysis of covariance ■ multivariate analysis of variance ■ nonparametric tests

In the late 1970s and the early 1980s, many journals including *Circulation Research* published articles warning about the misuse of statistical methods in the analysis of data.[1–7] Among the main errors identified were incorrect use of multiple comparisons of several independent groups, low statistical power, and inadequate analysis of repeated-measurement studies. The problem in the multiple comparisons of means from more than two populations has been well recognized since then, and now it is rare that misuse of repeated *t* tests is published in peer-reviewed journals. Although we still see manuscripts containing this problem, reviewers often prevent these articles with multiple *t* tests from being accepted and advise reanalysis to the authors. In place of the previous error, however, there is a tendency to overuse the Bonferroni correction for multiplicity. In contrast, the problem about repeated measurements on the same experimental units has not been completely resolved; inaccurate analysis is still frequent. Finally, inadequate power analysis often leads to incorrect acceptance of the null hypothesis of no difference between the groups. The problems may be partly perpetuated by the vagaries of statistical program packages; analysis of variance (ANOVA) with multiple-comparison tests is contained in almost all program packages for statistics, whereas the program for the analysis of repeated measures is not contained in usual packages for biomedical statistics, and power analysis is not included as often as it should be. Furthermore, although it is not difficult to analyze data with ANOVA (even by hand calculation), this is not so for repeated measures. In the early 1980s, the availability of computers was limited whereas electrical calculators became popular, but these were not well adapted to repeated-measures analysis.

In this review, several statistical methods are introduced that are useful to test various hypotheses but are not well recognized or sufficiently utilized. These methods are now easily available in some extended statistical program packages such as SPSS for the PC.[8] The questions indicated below are the target examples of this review. The aim of this review is not only to draw the attention of investigators to these tests but also to stress the conditions in which they are applicable. With the choice of appropriate statistical programs, the issue is no longer how to do the test but when to do it.

In this review, we will discuss the following topics:

1. What questions does the investigator need to ask about the distribution of the variables?
2. What do significance and probability value mean?
3. Type I and type II errors: how many measurements are needed? (power analysis)
4. What are our concerns about multiple comparisons?
5. Repeated measures: when and how should such analysis be used?
6. Is there a difference between two or more regression lines? (analysis of covariance)
7. How does one compare saturating functions, eg, Ca^{2+}-sensitivity curves of myofilaments, under multiple conditions? (MANOVA)
8. How is it appropriate to compare survival curves under different conditions? (actuarial life-table analysis and Kaplan-Meier analysis)
9. What are the factors that determine survival curves? (Cox regression model)

Original received July 23, 2002; revision received August 28, 2002; accepted September 3, 2002.

From the Institute for Clinical Research (H.K.), Osaka National Hospital, Osaka, Japan; Department of Pediatrics and Cardiovascular Research Institute (J.I.E.H.), University of California, San Francisco, Calif.

Correspondence to Hideo Kusuoka, MD, PhD, Director, Institute for Clinical Research, Osaka National Hospital, 2-1-14 Hoenzaka, Chuo Osaka 540-0006 Japan. E-mail kusuoka@onh.go.jp

Circulation Research **is available at http://www.circresaha.org** **DOI: 10.1161/01.RES.0000037427.73184.C1**

10. How does one test hypotheses in which the parameters show distributions far from the normal distribution? (nonparametric tests)
11. Is there a correlation between two parameters expressed by ordered category? (measures of correlation for R×C tables)
12. Do two indices evaluate the data similarly? (κ measure of agreement)
13. How much does a drug reduce a risk? (odds ratio and McNemar's test)
14. How does one predict risk in a patient? (logistic regression)

Basic Concepts in Statistics

Statistics has been defined as the art and science of dealing with the variability of measurements. Without variability, there would be no need for statistical analysis. If all people with essential hypertension had resting blood pressures of 150/100 mm Hg, then diagnosis would be easy. Furthermore, it would be simple to determine if a new agent lowered resting blood pressure. Give it to a few patients, and if the resting systolic blood pressure is measurably lower, we have established that the agent works. Unfortunately, in real life, there is variability, which is sometimes considerable. Even in a single patient, daily resting blood pressures vary, so that if the resting blood pressure was 150 mm Hg one day and 140 mm Hg on the next day after a drug had been given, we could not be sure that the decrease was due to the drug rather than to natural variability. In most statistical analyses, therefore, we need to compare a difference that might have been induced by a treatment or stimulus to some measure of variability: test statistic=difference/variability. The bigger the difference is relative to the variability, the more likely it is that the treatment caused the change. The differences we find depend on the problems being studied, but it is often possible to do things that reduce variability. Often the difference between an efficient and an inefficient statistical test depends on how we handle variability.

The numbers that we analyze statistically are conveniently classified into three groups (scales) that require different statistical tests.

Ratio scale: The ratio scale gives numerical data. Ratio numbers are the usual numbers that we deal with, where 4 is twice 2, and the interval from 2 to 4 is the same as the interval from 4 to 6. Ratio numbers are the subject of most parametric statistical tests like *t* tests, analysis of variance, and regression.

Ordinal scale: The ordinal scale gives qualitative but ordered data. Ordinal numbers are like +, ++, +++, ++++, or what seem to be ratio numbers but really stand for ordered categories; for example, grading symptom severity as 1 through 5. Grade 4 is not twice as severe as grade 2, and ++ is not twice +. Furthermore, the interval from grade 2 to 3 is not the same as the interval from grade 3 to 4. These numbers are analyzed by nonparametric tests such as Spearman's correlation coefficient or the Kolmogorov-Smirnov tests.

Nominal scale: The nominal scale gives categorical data. Counts in the categories (number alive versus dead or over and under some critical concentration) require analysis by Poisson or binomial statistics or χ^2 tests.

1. Distributions

Because *t* tests are the most commonly used tests, they will be used to illustrate some important issues. Like all statistical tests, they are based on a mathematical model. Like many parametric tests, the model requires the underlying distributions tested to be normal Gaussians. The *t* test is fairly robust, that is, it can tolerate some departure from normality without losing much efficiency, but large departures can be devastating. (Robustness means that the statistics work well for a wide variety of population types of the samples.) Easy ways of testing normality are to inspect the distribution after plotting a stem-and-leaf diagram,[9,10] noting that the standard deviation is about the same size as the mean (indicating severe rightward skewing), calculating skewness and kurtosis, or performing a Shapiro and Wilks test or the D'Agostino and Pearson test.[11,12] If the distributions are grossly abnormal, then either they need to be normalized by some transformation (square root, logarithmic, and reciprocal are the most frequently used[10]) or else one of the nonparametric tests needs to be used.

2. Significance of the Probability Value

Most investigators pay great attention to the probability value without really thinking what it is telling them. The probability value is the probability of a type I error, that is, the probability that the null hypothesis is true. If we have two groups, for example, a control and an experimental group, and the two means are different, the null hypothesis states that the treatment has not changed anything, and the observed differences could have come about by chance in drawing two groups from the same population. If this probability is high, we would not want to assert that the treatment has changed the outcome. If the probability is low, we might want to assert that the treatment has changed the outcome. Conventionally, this probability is set at ≤ 0.05 (also known as the type I or α error), but this is arbitrary. The great statistician Ronald Fisher originally suggested this probability because he regarded a 1 in 20 chance as being rare. Obviously, a 1 in 100 chance is even rarer, but he probably knew from experience that this would impose a standard that could seldom be met. However, the critical value of ≤ 0.05 is not an absolute requirement but merely a recommendation that will serve the purpose most of the time. For example, if by rejecting the null hypothesis we plan to start a new project that will cost millions of dollars and take years to complete, we might require a probability <0.01. If, on the other hand, we are screening for possible useful new treatments, a probability of <0.10 might be selected; after all, a 9:1 chance of being right is good odds. A statement like this is often seen in an article: "The treatment changed the flow from 5 mL $\cdot$ g^{-1} $\cdot$ min^{-1} to 3 mL $\cdot$ g^{-1} $\cdot$ min^{-1} ($P=0.07$), which is not significant, so that the treatment had no effect." This cannot be right. Flow decreased considerably, but we are not quite as confident in rejecting the null hypothesis as if $P=0.04$. Nevertheless, it would be inefficient to discard the possibility that the treatment had caused flow to decrease; at the very least, we might want to do more experiments.

The other important aspect of the probability value is the term "significance." In statistics, this has the specific mean-

ing of rejecting the null hypothesis at a given probability. It has nothing to do with the meaning of significance in ordinary language, where it implies important or noteworthy. If, for example, we do a clinical trial with two antihypertensive drugs and recruit 100 000 people to take each drug, it might turn out that drug A lowered blood pressure by 1 mm Hg more than drug B, and that this difference was statistically significant. This small difference would almost certainly be of no clinical or physiological importance. One way of emphasizing these absolute quantities is to give the probability value and to calculate the confidence limits for the difference obtained.[13]

Most statistical tests involve examining a difference between two or more groups and then comparing that difference to some measure of variability. The most efficient way to evaluate such a test is to look at the difference in absolute units and decide if it is physiologically or clinically important. If it is too small to be of interest, then whether it is statistically significant or not is of no particular value. If the difference is large enough to be meaningful, then it needs to be related to the measure of variability. If the ratio is large, it implies that the difference may be statistically significant, that is, that the null hypothesis can be rejected. If the ratio is small, then we have to ask ourselves why. The measure of variability can be inflated because the distributions are not normal or do not have similar standard deviations, and if that is the reason, then either the data must be transformed to fit the requirements of the test or else a nonparametric or distribution-free test must be used. On the other hand, the measure of variability may be perfectly adequate, but the sample size may be too small. This leads to the important concept of statistical power (see below).

Reviewers often make an irritating minor error. The *t* test yields a significant result, but the reviewer comments that because the distribution is grossly abnormal, a *t* test should not be done. If the distribution is very abnormal, that may cause a real difference to be regarded as not significant because the variability will have been inflated by the skewed distribution. On the other hand, if despite the abnormal distribution the *t* test comes out to be significant, that is never a false result. Had the distribution been normalized, then the *t* test would have been even more significant, but that is not usually required once significance has been demonstrated.

3. Type I and Type II Errors and Power Analysis

Statistical power is the probability of getting a statistically significant result if there is a biologically real effect in the population being studied. The type I error mentioned above is the probability of rejecting the null hypothesis falsely. Its counterpart is the type II error (termed β), the probability of accepting the null hypothesis falsely, that is, of rejecting the fact that there is a difference between the two groups. The power of a test is calculated as 1-β, a measure of the ability to detect a real difference if it is there. If the sample size is too small, then it may not be possible to establish the significance of a given difference, but that does not mean that the difference is not there. Power analysis allows us to be certain that we have looked hard enough for the difference. Most measures of variability use some form of standard error, which for a simple *t* test is the standard deviation divided by the square root of *n*, the number in the sample. As the sample size increases, the standard deviation hovers about its true value, but the standard error decreases progressively.

One of the first studies to draw attention to this problem in medicine was that by Freiman et al,[14] who examined 71 randomized trials that compared the effects of two drugs or treatments and concluded that there was no statistically significant difference between their effects. They showed that, in many of those studies, the actual responses were quite large, but because the sample sizes were too small there was a greater than 10% chance of missing a true 25% therapeutic improvement in 67 of the trials and a true 50% therapeutic improvement in 50 trials. In some instances, this led the investigators to discontinue studying the new treatment and to conclude that it was of no benefit. Clearly this is an undesirable outcome; a 25% improvement in the cure rate in any disease would be very welcome.

Unfortunately, this important article was largely ignored, as documented by inadequate power found in many medical studies by Reed and Slaichert[15] and more recently in many articles published in the *American Journal of Physiology* as reviewed by Williams et al.[7] The error is pervasive. We recommend that investigators take this issue seriously. It is both inefficient and probably also unethical for investigators, clinical or other, to plan and execute an experiment that cannot answer the question.

There is a vast literature on this subject, and much of it is accessible to the average investigator. The concepts and the necessary tables are readily available in several standard books[16–18] and discussed in many articles. There are also several statistical programs for power analysis that are either free or for purchase that have been evaluated by Thomas and Krebs,[19] and most of these can be found on an excellent Web site *Web Resources on Power Analysis*.[20] The calculations are best done a priori, that is, in planning the study and before starting it, but they can also be done post hoc in determining the power of a study that has been completed. The principle of the a priori calculation is simple and will be illustrated for the unpaired *t* test. The first thing is to decide what a meaningful difference between the two groups would be. Is it a 25% reduction in mortality? Is it a 40% reduction in interleukin-6 concentration? A 20 mm Hg decrease in diastolic blood pressure? Then an estimate of the variability of the data, as typified by the standard deviation, is needed. This is usually available from similar studies in that field but occasionally may have to be obtained from a pilot study. The ratio of the absolute difference to the standard deviation is symbolized by *d* or δ, often known as the effect size. Next, set the values for α and β. The type I error α is conventionally set at 0.05, but if this results in unattainable numbers, a value of 0.10 could readily be used. The type II error β is optimally 0.05 (power 0.95), but once again if this leads to unacceptable high numbers, then β could be increased to as high as 0.20 (power 0.80). With these values for α, β, and *d*, either the tables in the books can be used, or the values can be entered into one of the computer programs (self-standing or online) to determine what number *n* is needed for each group. The post

TABLE 1. Multiple Comparisons

Number of Pairs (N)	Probability of Correctly Rejecting the Null Hypothesis	Probability of Falsely Accepting the Null Hypothesis (I)	Ratio I/N
1	1–0.05=0.95	0.05	0.050
2	0.95^2=0.9025	1–0.9025=0.0975	0.049
3	0.95^3=0.8574	1–0.8574=0.1426	0.047
4	0.95^4=0.8145	1–0.8145=0.1855	0.046
5	0.95^5=0.7738	1–0.7738=0.2262	0.045
6	0.95^6=0.7351	1–0.7351=0.2649	0.044
7	0.95^7=0.6983	1–0.6983=0.3017	0.043
8	0.95^8=0.6634	1–0.6634=0.3366	0.042
9	0.95^9=0.6302	1–0.6302=0.3698	0.041
10	0.95^{10}=0.5987	1–0.5987=0.4013	0.040

hoc calculation is used to find out the power of a completed study. It is done in the same way, except that *d* and *n* are known, and α is usually 0.05. If the power of the test turns out to be low (eg, 0.4), then there is no way to tell if the two groups are or are not different.

4. Some Concerns About Multiple Comparisons

Glantz[1] identified these as among the most frequent errors of statistical analysis. Wallenstein et al[6] dealt very effectively with the problem, but it appears to us that the pendulum has swung too far in the other direction, that is, that correction for multiplicity is sometimes used when it is not needed. People have a great deal of difficulty in deciding when corrections for multiplicity are needed, and there are even times when statisticians disagree.[21] Nevertheless, the general principles are straightforward. In addition, we would like to describe some other analyses that can be used in certain circumstances when the issues of multiple comparisons arise.

Let us illustrate that repeated *t* tests shift the probability from a single test. For a simple example, consider that you are trying to decide whether to go home early or stay and work for another 2 hours. To make the decision you will toss a coin. If it lands with heads up, you will go home early; if not, you will stay and work. You toss the coin, and it comes up tails. You toss it again, and again it comes up tails. You continue tossing until it comes up heads, so you pack up and go home early. Obviously, at the first toss, there is a 50:50 chance of heads coming up, but as you continue tossing, there will eventually be near certainty that a head will appear; the chances of getting 10 tails in a row are 0.0009765625.

For a more detailed discussion, we can do no better than use an explanation given by Tukey.[22] Draw two groups at random from a population with a normal distribution. Set the probability of falsely rejecting the null hypothesis (which we know to be true) at 0.05. Therefore, the probability of correctly accepting the null hypothesis is 1−0.05=0.95. Now draw two more groups at random from the same population, and once again there is a probability of 0.05 of falsely rejecting the null hypothesis and 0.95 of correctly accepting the null hypothesis. Now what happens if we state that we will reject the null hypothesis if either of the two sets shows a significant difference? The probability of correctly accepting the null hypothesis for both sets is the product of the two probabilities: 0.95×0.95=0.9025. Therefore, the probability of falsely rejecting the null hypothesis is 1−0.9025=0.0975. In other words, by giving ourselves two chances to reject the null hypothesis, we have almost doubled the chances of falsely rejecting it. If we continue to draw pairs of groups at random from the parent population, the risk of falsely rejecting the null hypothesis increases steadily, as shown in Table 1.

This table shows that, as the number of *t* tests increases, the risk of a type I error increases, even though for each individual *t* test the risk remains at 0.05. One of the ways of reducing the type I error is to divide the probability of making a type I error by the number of comparisons (*t* tests). As shown in the last column, this ratio remains close to the conventional 0.05 value. This is the basis of the Bonferroni correction.

The issue to be determined is when to apply this correction. It is essential to realize that there are two different ways of comparing data, and this can be exemplified by the study of Creasy et al.[23] They produced growth retardation in fetal lambs by embolizing the placenta with 15-μm-diameter microspheres and then compared control and growth-retarded lambs for weights of 9 different organs as well as for arterial oxygen and carbon dioxide tensions. In all, there were 11 comparisons. We now introduce the conventional definitions of two error rates, the comparison-wise error rate and the family-wise (or experiment-wise) error rate[21]:

Comparison-wise error rate=number of comparisons leading to rejection of the null hypothesis/total number of comparisons

Family-wise error rate=number of families leading to rejection of the null hypothesis/total number of families

The comparison-wise error rate is the familiar type I error. Each organ weight, for example, can be validly compared by *t* test without any need for correction, and the conventional 0.05 value can be used (Figure 1).

On the other hand, if the unit of comparison is the whole family of comparisons, and if any of the 11 separate comparisons leads to the conclusion that placental embolization affects the fetus, then we need protection against the inflated

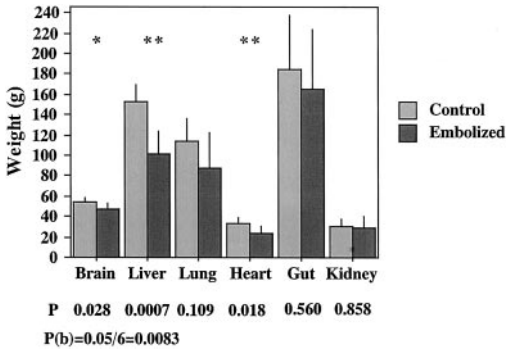

Figure 1. Weights of 6 organs (Creasy et al[23]). Short vertical lines are standard deviations. *$P<0.05$; **$P<0.01$. P(b) indicates critical value with Bonferroni correction.

error due to multiplicity. Failure to distinguish these two types of errors leads many investigators to use the Bonferroni (or other) correction when it is the comparison-wise error rate that is needed. By applying the Bonferroni correction unnecessarily, the value of α for the type I error is made so small as to be difficult to attain. In Figure 1, in which the data for 6 organs are shown, three of the organs show significant differences with a simple *t* test, but if the Bonferroni correction has been made, only one of those would have shown a significant difference. If, however, 50 separate comparisons had been made, then the Bonferroni correction would have required $P<0.05/50=0.001$, and probably no single comparison would have been significant. For the group as a whole, an analysis of variance correctly identifies a difference due to embolization, with $P=0.028$.

There is another way in which multiple time or dose comparisons can be made without involving the Bonferroni adjustment. Consider an experiment in which interleukin-6 (IL-6) is measured every 30 minutes for 3 hours in control animals and in treated animals to which a potential agonist has been given. Excluding the control values at time zero, there are 6 time points and 6 possible comparisons. It could easily happen that, if the number of animals is small and the differences due to the agonist are small, no significant differences can be established at any time point, yet it is clear from the figure that the agonist consistently increased IL-6 concentrations (Figure 2).

Figure 2 shows control concentrations (curve A) and two sets of agonist-stimulated concentrations (curves B and C).

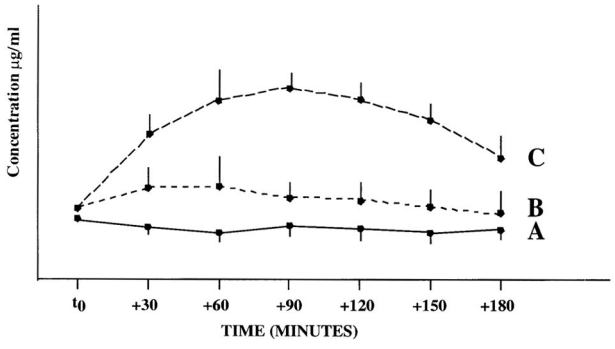

Figure 2. Measurement of IL-6 under different conditions. The short vertical lines are standard errors.

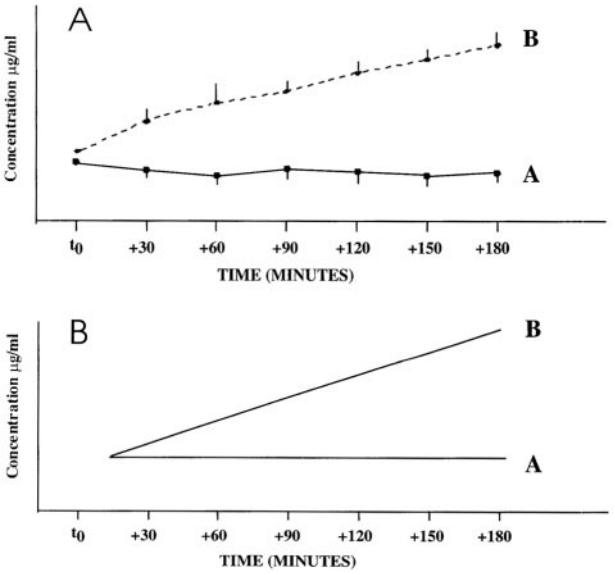

Figure 3. Example in which the concentrations linearly increase with time. The short vertical lines in panel A are standard errors.

Curve C shows data in which each time point after time zero has high concentrations. If the concern is to show that curve C is different from the control A, one could either do an analysis of variance or else just examine the area under the whole curve. If this area is calculated for each subject, then it can be compared with the null hypothesis (ie, that there is no change) by a simple *t* test. By inspection, curve C is clearly different from curve A, but curve B with lower concentrations might not be shown to be significantly different without testing. If, in comparing curves A and B, we did multiple *t* tests and used a Bonferroni correction, it is possible that none of the differences at the different time periods might be significant, yet the curve as a whole would show a significant difference.

There is one other format that needs discussion. Consider Figure 3A, where the experimental group has an approximately linear increase with time.

The way to handle this depends on what question is being asked. If the question is "Has the agonist changed the concentrations?," then the most efficient way to test significance is to fit a straight line to each curve by linear regression and then compare the slopes by analysis of covariance. If the control slope is zero, it might be sufficient to show that the experimental line has a slope that is significantly different from zero. This method of testing is much more sensitive than doing an analysis of variance, because by using knowledge of the relation between the x and y variates, the residual variability is greatly reduced. There is, however, another question, namely: "When does the first increase in concentration occur?" There is a simple answer to this but also a caveat. The simple answer is to do *t* tests at each time point, and the first one that is significant indicates a significant departure from control. However, this answer applies only to the data under consideration and cannot be generally applied. If, for example, we think of a chemical reaction in which the product increases linearly with time (Figure 3B), that increase may start as soon as the chemicals are mixed. The fact that an

early time point cannot be shown to be different from control might merely mean that the numbers of experiments are too small to show a significant difference because the power of the test to show a small difference at that time is low. If we believe that it is important to show that the reaction begins, for example, 30 seconds after the chemicals are mixed, then we need to design an experiment with sufficient power to determine a small difference.

5. Repeated-Measures Analysis of Variance

Wallenstein et al[6] discussed how to handle this type of analysis in detail, but we would like to emphasize the reasons for these more complex procedures. Consider an experiment that takes two groups of subjects (people or animals), measures baseline concentrations of IL-6, and then measures IL-6 concentrations again after giving either intravenous saline or else an equivalent volume of a putative inhibitor of IL-6 production. The hypothesis is that the inhibitor will decrease the concentration of IL-6. This experiment can be done in two ways. In one, each subject provides one pair of measurements, and we do an unpaired *t* test to determine if the differences in the control and treated groups are the same or not. The other way would be to use several subjects in each group but to give each subject increasing doses of the putative inhibitor. This method has the advantages of getting more measurements out of a given number of subjects and of allowing determination of a dose-response relationship. It has the disadvantage of blurring the distinction between the groups, because the variability from one subject to the next might be great enough to conceal differences between the groups. Repeated measurements are usually made at sequential times or doses. Whenever multiple measurements are made on the same subject (patient or animal) or object (eg, cell culture, blood sample, or vascular ring), special procedures to deal with repeated-measures analysis are needed.[24]

As a specific example, an experiment is done to test the accuracy of a skin electrode for measuring arterial oxygen tension (Pa_{O_2}) in subjects with and without tissue edema. With an arterial needle or catheter in place, a number of simultaneous measurements are made of arterial tension in blood (Pa_{O_2}) and via the skin electrode ($tcPa_{O_2}$) in each of several healthy subjects and patients with tissue edema over as wide a range of arterial oxygen tensions as possible. When the results are displayed as an x-y plot, the relationship within each group is linear, with a similar slope in each patient but with slight differences in the intercepts. An idealized result is given in Figure 4A for the healthy group.

How should this experiment be analyzed, other than to report each individual relationship? One thing that should not be done is to pool all the points for the healthy subjects into one group, the points for the edematous subjects in another group, and then make a simple comparison. If that is done, two errors are likely to occur. One is that the slope of each x-y relationship is likely to be incorrect, as shown in Figure 4B. In fact, Glantz and Slinker[25] gave a striking example in which each of three subjects had no relation between x and y (horizontal line on the graph), but pooling the data led to the creation of a significant slope. The second error is that the estimate of variability is likely to be greatly increased

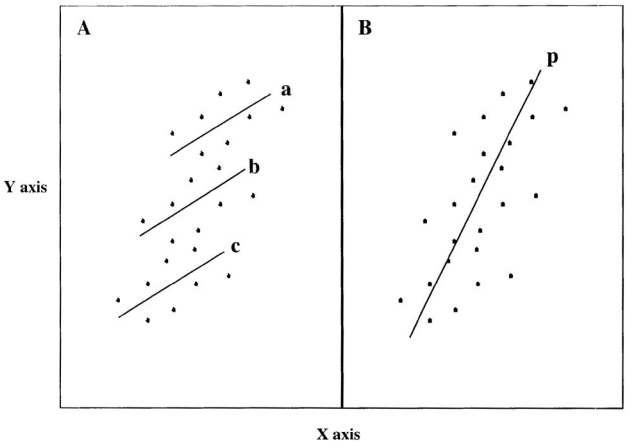

Figure 4. Principle of repeated-measures analysis. Left, Hypothetical relation between Pa_{O_2} (x-axis) and tcP_{O_2} (y-axis) for 3 subjects a, b, and c in the same group; all three slopes are the same, but the intercepts are different. Right, Three sets of points pooled (p) into a single group, which has a slope that differs from the individual slopes and also has a greater standard deviation from regression than each individual data set has. Repeated-measures analysis avoids these sources of error.

because the intrinsic variability of the points about each line will be added to the variability of the lines from each other. If we pool the data for the group with edema, we would then compare two sets of data with incorrect slopes and exaggerated variability. Even if the slopes were not misleading, the increased variability would make it more difficult to show differences between the two groups.

The way to do the analysis is to partition the total variability into the differences between the two groups—between the subjects within each group, that is, the differences in the y values (here $tcPa_{O_2}$) due to differences in the x value (here Pa_{O_2}), and the residual differences. The component of variability due to difference between subjects is combined with the residual variability in Figure 4B, but with special techniques it can be removed, so that the residual variability is reduced to the average of the three individual sets shown in Figure 4A.

To give a more detailed example, a pulse rate was measured every 10 minutes for 4 times in patients with or without a drug. The data are arranged with four variables, so that four pulse rates are recorded for each patient. A factor that encompasses each set of repeated measurements is defined as a within-subjects one. In this example, time is defined as a within-subjects factor, and drug is specified as a between-subjects factor because it divides the groups of subjects into two. The hypotheses about the effects of both the between-subjects factors and the within-subjects factors can be tested by repeated-measures analysis. The interactions between factors can be checked as well as the effects of individual factors.

Measurements of more than one variable for the different levels of the within-subjects factors can be analyzed by a doubly multivariate repeated-measures analysis. The extension of the above example, that is, the measurement of pulse and respiration at 4 different times on each subject, is an example of doubly multivariate repeated measures.

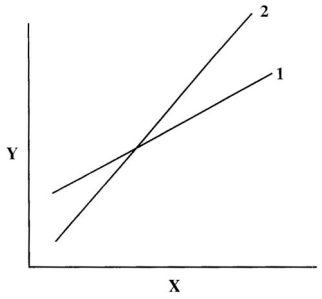

Figure 5. Example for ANCOVA. At low values of x, the y values in group 1 exceed those in group 2, but the reverse is true at high values of x.

6. Analysis of Covariance (ANCOVA)

This is the test to determine whether two or more regression lines obtained from the data are identical or not. The analysis of covariance has the combined features of analysis of variance (ANOVA) and regression. In ANOVA, the typical model for the value of the observation consists of the population means of different classes and the residual. In contrast, the model in ANCOVA consists of one more variable that is linearly related to the observed value. That is, a given value is a sum of the population mean, the contribution of regression, and the residual.

ANCOVA is useful to study regressions in multiple classifications. For example, the relation between variables x and y in the groups with or without a treatment is investigated, and the relation is supposed to be linear. To test the hypothesis whether the treatment shifts the regression line between x and y, ANCOVA should be used. Investigators should be aware that ANCOVA produces two results. First, it tests to find out if the two slopes are significantly different. If they are not, then an average slope is calculated and the test determines if one line lies significantly higher or lower than another. (There can be more than two groups.) Thus, in Figure 4, the three linear regressions in the left panel can be compared and, if the slopes are not significantly different, they can be combined into a single average slope for that group. If the question is of interest, the positions of the different slopes can be compared. If the two slopes are significantly different, then it may not make sense to ask if one line lies above another because of the dependence of height on the value of the x variate (Figure 5).

Another application of ANCOVA is adjusting bias in studies. For example, in a study in which the responses of some drugs are measured in rats, if it is known that the response is related to the body weight, the differences in the responses may not be due to the drug effects but to the different weights among the groups. Thus, it is necessary to correct the bias that comes from weight. When the relation of the response and the weight can be assumed to be linear, ANCOVA is applicable.

7. Multivariate Analysis of Variance (MANOVA)

It is important to differentiate the behaviors of the heart under different conditions. Usually, the behavior is characterized by its calcium responsiveness. Measurements of calcium responsiveness in hearts, muscles, or myocytes under different conditions produce sigmoid, saturating data for contractile activation. However, the proper statistical technique to detect the difference in calcium responsiveness is not well established. It is not difficult to test for changes in the maximal responses, but it is not easy to test the sensitivity (ie, the steepness and the midpoint of the curves). The statistical test of EC_{50}, that is, the calcium concentration that gives the 50% of maximum response, is often used, but the results depend greatly on how to calculate EC_{50} and also on the maximum value.

When the external calcium concentration is fixed, MANOVA is applicable. MANOVA considers the effects of factors on several dependent variables at once, using a general linear model. The factors divide the cases into groups. The hypotheses tested are similar to those in univariate analysis (ANOVA), except that in multivariate analysis, a vector of means replaces the individual means. For example, the developed pressure was measured at 5 different calcium concentrations of the perfusate in the control and stunned myocardium (see Figure 5 in Kusuoka et al[26]). The data in each heart were expressed as a vector whose elements were the developed pressure at each calcium concentration. MANOVA indicated a statistically significant difference in calcium responsiveness between the control and the stunned heart.

In some experiments, it is not possible to control the calcium concentration, or different sets of calcium concentrations are used among the groups. MANOVA is then not applicable. As an alternative method, the relation is fitted by some equation, and MANOVA is applied to the set of the parameters that characterizes the equation. For example, the relation between calcium concentration ($[Ca^{2+}]$) and tension (T) can be fitted with Hill's equation, for example: $T=K[Ca^{2+}]^n/(1+K[Ca^{2+}]^n)$. Then this relation is changed into a linear form: $\log\{T/(1-T)\}=\log K+n\log[Ca^{2+}]$.

In this example, ANCOVA may be used to test the difference among the linear regression forms of Hill's equation. MANOVA can also be used to test the set of Hill's constants (K) and Hill's coefficients (n). These methods are widely applicable to test the dose-response curves under different conditions.

8. Nonparametric Tests

Most classical statistics have two major assumptions. One is that the data come from a specific distribution, usually the normal distribution. Statistics are calculated based on the distribution parameters such as the mean and variance. Another assumption is that the groups being analyzed have equal variances. Although most statistics are robust enough to withstand minor violations of these assumptions, it may happen that the data violate these assumptions so substantially that the results may not be reliable.

There are many different types of distributions, and robust techniques have been developed to handle them. Nonparametric tests can analyze data under unfavorable circumstances; they are also known as distribution-free tests because there is no need for assumptions about the distribution. Nonparametric tests have another advantage in that they are

TABLE 2. Parametric and Nonparametric Statistical Tests of Hypothesis

I. Tests for 2 independent samples
(1) Categorical data: Fisher's test, χ^2 test
(2) Ordered data: Mann-Whitney test, Wilcoxon test
(3) Numerical data: Mann-Whitney test (N), Wilcoxon test (N), Student's *t* test (P)
II. Tests for 2 related samples
(1) Categorical data: Fisher's test, χ^2 test
(2) Ordered data: sign test, Wilcoxon's signed rank test
(3) Numerical data: Wilcoxon's signed rank test (N), sign test (N), paired *t* test (P)
III. Tests for more than 2 independent samples
(1) Categorical data: χ^2 test
(2) Ordered data: Kruskal-Wallis test
(3) Numerical data: Kruskal-Wallis test (N), one-way ANOVA (P)
IV. Tests for more than 2 related samples
(2) Ordered data: Friedman test
(3) Numerical data: Friedman test (N), one-way repeated-measures ANOVA (P)
V. Correlation
(2) Ordered data: Spearman correlation coefficient
(3) Numerical data: Spearman correlation coefficient (N), Pearson correlation coefficient (P)

(N) and (P) in the methods for numerical data indicate nonparametric and parametric methods, respectively.

resistant to gross mistakes in measuring, recording, or copying data for statistical analysis, because they are little affected by gross errors in a few observations.

For some tests, both parametric and nonparametric versions exist. Generally, the parametric one is more sensitive than the nonparametric one and should be used when the necessary assumptions are met. However, nonparametric tests, although being distribution-free, are not assumption-free. These tests require certain assumptions, but in general they are satisfied more easily than those required by parametric ones. The relation between the parametric and the nonparametric methods is summarized in Table 2. Typical nonparametric tests are the Mann-Whitney *U* test and the Wilcoxon test that can replace the unpaired and paired *t* tests, respectively. There are, however, many other nonparametric tests to fit almost any contingency.[27,28]

9. Measures of Correlation for R×C Tables

This is a test to determine whether there is a correlation between two variables with ordered categories. For example, the degrees of cell damage in an assay are categorized as severe, moderate, mild, or none by two different measures (eg, morphology and contractile activation). The correlation is evaluated by the Spearman correlation coefficient, the Pearson correlation, and the linear-by-linear association χ^2.[29] For the Spearman correlation coefficient, the rank order of each value is used to compute the Pearson correlation. The linear-by-linear association χ^2 is simply the square of the usual Pearson correlation multiplied by the sample size minus 1. For the Pearson correlation, it is assumed that the data come from a normal distribution, and this is not satisfied in two-way table data. In contrast, the Spearman correlation requires no assumption about the nature of the population sampled.

10. Life-Table Analysis and Kaplan-Meier Method

In some studies, it is necessary to examine the time to occurrence of a critical event of interest such as survival time of animals or patients after one of several interventions or no intervention at all. Subjects can enter a study at various times. The time is measured from the start of observation, ie, the time that the subject enters the study, until the event is observed. (The event need not necessarily be adverse.) However, these data usually include some subjects for whom the second event is not observed, for instance, the subjects are still alive at the end of study or the subjects are lost to follow-up. These subjects are termed "censored," and they make this kind of study inappropriate for the application of traditional statistical methods.

Life-table analysis is useful for processing this type of data. In life-table analysis, the period of observation is divided into smaller time intervals. For each interval, the probability of the terminal event for each interval is calculated from the subjects who have been observed at least until that period; the probability is given by dividing the number of subjects experiencing the terminal event during the interval by the number of subjects entering the interval alive. Then, the probabilities estimated from each interval are used to estimate the overall probability for the event to occur at different time points.[30,31] There are two similar ways of constructing these tables.[32,33] Thus, when survival times have been categorized into time intervals such as days, months, or years, only actuarial life-table analysis is applicable. If exact times of the events are known, more precise estimates are available by the Kaplan-Meier method. The probability of a terminal event is calculated at every occurrence of the event. This makes Kaplan-Meier techniques useful for studies with few subjects where the survival intervals are variable. The excellent review articles by Peto et al[34,35] give more details about this technique.

11. Cox Regression Analysis

Cox regression, like life-table analysis and Kaplan-Meier survival analysis, is a method for modeling time-to-event data in the presence of censored subjects. However, Cox regression is different from others because this method allows the inclusion of predictor variables (covariates) in the models. For example, a Cox regression model with cigarette usage and gender as covariates is constructed to test the hypothesis regarding the effects of gender and cigarette usage on time-to-onset for lung cancer. Cox regression can handle the censored subjects correctly, and it provides estimated coefficients for each of the covariates. The impact of multiple covariates can then be assessed in the same model. More generally, Cox regression can allow for differences in the baseline characteristics of the groups that are being compared, whether in a randomized or a nonrandomized trial.

The Cox proportional hazards regressions model is popular in part because it requires fewer assumptions than some other

TABLE 3. Odds Ratio

	Alive	Dead	Subtotal
A	20	8	28
B	12	12	24
Subtotal	32	20	52

For an example, 52 patients were divided into 2 groups, A and B, by some factor such as smoking, and the prognosis was evaluated 5 years later. The odds ratio is (20×12)/(12×8)=2.5. This is interpreted as the odds of being alive in group A are 2.5 times as great as being alive in group B.

survival models. However, it should be used only if the assumption of proportional hazards is fulfilled. This means that the hazards for individuals with different covariates are constant over time. For example, the differences between the death rates for two different surgical procedures should be roughly constant at different times after the procedure. If the survival curves cross, then this assumption is violated, and an extended version of the Cox model must be used, perhaps by dividing the analysis into localized time periods. Another assumption that needs to be tested is that the covariates have a multiplicative effect on the hazard rate. For example, if heavy smokers have a risk of lung cancer three times that of light smokers and men have a risk that is twice that of women, then the risk for men who are heavy smokers should be about 2×3=6 times that of women who are light smokers. Failure to meet this criterion points out the need for special techniques.

12. Kappa (κ) Measure of Agreement

This is the type of test to determine whether observer A and observer B similarly evaluate the events of interest. The data are presented in square tables of dimensions R×R; the number of rows and columns is the same because each subject is classified twice.

Kappa is a measure of interrater agreement that assesses if the counts in the diagonal cells (the subjects who receive the same rating) differ from those expected by chance. When all off-diagonal cells are empty, κ achieves its maximum value, 1.0. Kappa is judged by using asymptotic standard error to construct a t statistic to test whether the measure differs from 0. Values of κ greater than 0.75 indicate excellent agreement beyond chance, values from 0.40 to 0.75 indicate fair to good, and values below 0.40 indicate poor agreement.

Another application of the κ measure is the test for coexistence of two phenomena in the same subjects. For example, the degrees of systolic and diastolic dysfunction are classified in 4 categories, and the patients are diagnosed for systolic and diastolic functions. If systolic and diastolic functions are always altered in the same manner, κ should indicate excellent agreement.

13. Odds Ratio and McNemar's Test

There are similarities between the t test and the χ^2 test. Both of them, if significance is reached, allow rejection of the null hypothesis that for the t test is the equality of means and for the χ^2 test is the equality of proportions. If the null hypothesis is rejected, then the magnitude of the difference in the t test is just the difference between the two means, whereas for the χ^2 test it is the odds (cross-product) ratio (Table 3); for both of these differences confidence limits can be set. One minor point about the χ^2 test with a 2×2 table is whether or not to make a Yates correction, that is, to reduce the absolute magnitude of the difference between observed and expected values by 0.5. Most authorities recommend this correction (which makes the test more conservative), but there are those who disagree. The issue can be avoided entirely by doing Fisher's exact test that, with current computer programs, can be done with almost any sample size.

TABLE 4. McNemar's Test

	Placebo	
Indomethacin	Closed	Open
Closed	65	27
Open	13	40

Another similarity is that both tests come in unpaired and paired versions. The usual χ^2 test is an unpaired test, and its paired counterpart is the McNemar's test. In the study summarized in Table 4, pairs of premature infants with a patent ductus arteriosus were matched for gestational age and sex; one of each pair was selected at random to receive indomethacin and the other to receive a placebo. Then, the subjects were evaluated for closure of the ductus. The numbers in the 2×2 table show the outcomes for each pair. In 65 pairs, both infants closed the ductus, and in 40 pairs the ductus remained open in both members of the pair. These concordant results give no information about the effect of treatment. The other two cells show discordance; in 27 pairs, one member closed the ductus with indomethacin but not with placebo, and in 13 pairs, one member closed the ductus with placebo but not indomethacin. The null hypothesis that indomethacin has no greater effect than placebo would give 20 pairs in each of these discordant cells and is tested by doing a χ^2 test on the two discordant cells. Note that to do a standard χ^2 test on all 4 cells would be meaningless.

14. Logistic Regression

In the usual linear regression, $y=c+bx$, we determine the factors (x_1, x_2, etc) that explain the variability in y, a continuous-response variable; for example, height (y) as a function of age (x) or glucose concentration (y) as a function of insulin dose (x). Sometimes, however, the response variable y is dichotomous, being either a success or a failure; for example, survival at 30 days of age for a premature infant related to birth weight or gestational age. To determine the probability (p) that an infant of a given birth weight survives to 30 days, we use logistic regression: $\ln\{p/(1-p)\}=c+bx$. Glantz and Slinker[25] and Pagano and Gauvreau[36] describe the principles and problems of this technique well.

Conclusions

In summary, we have introduced a number of different statistical methods that are not always familiar to basic and clinical cardiologists but may be useful for revealing the correct answer from the data. These methods are now generally available in statistical program packages. Research-

ers need not know how to calculate the statistics from the data but are required to select the correct method from the menu and interpret the statistical results accurately. We hope that this review promotes more suitable application of statistical analysis in *Circulation Research*.

Acknowledgments

This work was supported in part by Program Project Grant HL25847 from the National Heart, Lung, and Blood Institute of the NIH (to J.I.E.H.).

References

1. Glantz SA. Biostatistics: how to detect, correct, and prevent errors in the medical literature. *Circulation*. 1980;61:1–7.
2. Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J*. 1977;1:85–87.
3. Harrell FE, Lee KL, Grizzle JE. Additional comments on "some statistical methods useful in circulation research." *Circ Res*. 1981;48:592–594. Letter.
4. Levine HD. Comments on "some statistical methods useful in circulation research." *Circ Res*. 1981;48:592. Letter.
5. Rosen MR, Hoffman BF. Statistics, biomedical scientists, and circulation research. *Circ Res*. 1978;42:739. Editorial.
6. Wallenstein S, Zucker CL, Fleiss JL. Some statistical methods useful in circulation research. *Circ Res*. 1980;47:1–9.
7. Williams JL, Hathaway CA, Kloster KL, Layne BH. Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am J Physiol*. 1997;273:H487–H493.
8. SPSS Inc. *SPSS Application Guide*. Version 8. Chicago, Ill; 1998.
9. Tukey JW. *Exploratory Data Analysis*. Menlo Park, Calif: Addison-Wesley Publishing Co; 1977.
10. Velleman PF, Hoaglin DC. *Applications, Basics and Computing of Exploratory Data Analysis*. Boston, Mass: Duxbury Press; 1981.
11. Madansky A. *Prescriptions for Working Statisticians*. New York, NY: Springer-Verlag; 1988.
12. Zar JH. *Biostatistical Analysis*. 3rd ed. Upper Saddle River, NJ: Prentice Hall; 1996.
13. Gardner MJ, Altman DG. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines*. London, UK: BMJ Publishing Group; 1995.
14. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR The importance of β, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med*. 1978;299:690–694.
15. Reed JF III, Slaichert W. Statistical proof in inconclusive 'negative' trials. *Arch Intern Med*. 1981;141:1307–1310.
16. Kraemer HC, Thiemann S. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, Calif: Sage Publications, Inc; 1987.
17. Cohen J. *Statistical Power Analysis for Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
18. Lipsey MW. *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, Calif: Sage Publications, Inc; 1990.
19. Thomas L, Krebs CL. A review of statistical power analysis software. *Bull Ecol Soc Am*. 1997;78:126–139.
20. US Geological Survey, Patuxent Wildlife Research Center. Web resources on power analysis. Available at: http://www.mp1-pwrc.usgs.gov/powcase/powlinks.html. Accessed September 12, 2002.
21. Dunnett CW. Multiple comparisons. In: McArthur JW, Colton T, eds. *Statistics in Endocrinology*. Cambridge, Mass: MIT Press; 1970:79–103.
22. Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science*. 1977;198:679–684.
23. Creasy RK, Barrett CT, de Swiet M, Kahanpaa KV, Rudolph AM. Experimental intrauterine growth retardation in the sheep. *Am J Obstet Gynecol*. 1972;112:566–573.
24. Ludbrook J. Repeated measurements and multiple comparisons in cardiovascular research. *Cardiovasc Res*. 1994;28:303–311.
25. Glantz SA, Slinker BK. *Primer of Applied Regression and Analysis of Variance*. New York, NY: McGraw-Hill, Inc; 1990.
26. Kusuoka H, Porterfield JK, Weisman HL, Weisfeldt ML, Marban E. Pathophysiology and pathogenesis of stunned myocardium: depressed Ca^{2+} activation of contraction as a consequence of reperfusion-induced cellular calcium overload in ferret heart. *J Clin Invest*. 1987;79:950–961.
27. Conover WJ. *Practical Nonparametric Statistics*. 2nd ed. New York, NY: John Wiley & Sons; 1980.
28. Siegal S, Castellan NJ Jr. *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York, NY: McGraw-Hill; 1988.
29. Snedecor GW, Cochran WG. *Statistical Methods*. 8th ed. Ames, Iowa: Iowa State University Press; 1989:193–195.
30. Grunkemeier GL, Starr A. Actuarial analysis of surgical results: rationale and method. *Ann Thorac Surg*. 1977;24:404–408.
31. Grunkemeier GL, Wu Y. Actual versus actuarial event-free percentages. *Ann Thorac Surg*. 2001;72:677–678.
32. Muenz LR. Comparing survival distributions: a review for nonstatisticians, I. *Cancer Invest*. 1983;1:455–466.
33. Muenz LR. Comparing survival distributions: a review for nonstatisticians, II. *Cancer Invest*. 1983;1:537–545.
34. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient, I: introduction and design. *Br J Cancer*. 1976;34:585–612.
35. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient, II: analysis and examples. *Br J Cancer*. 1977;35:1–39.
36. Pagano M, Gauvreau K. *Principles of Biostatistics*. Belmont, Calif: Duxbury Press; 1993.